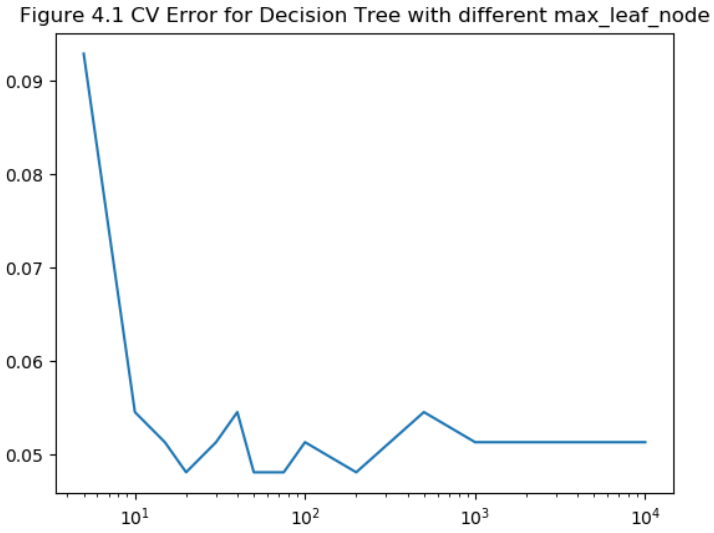
Homework 4: Decision Trees + Analysis

CS412

Due: April 23rd, 11:30pm on Gradescope

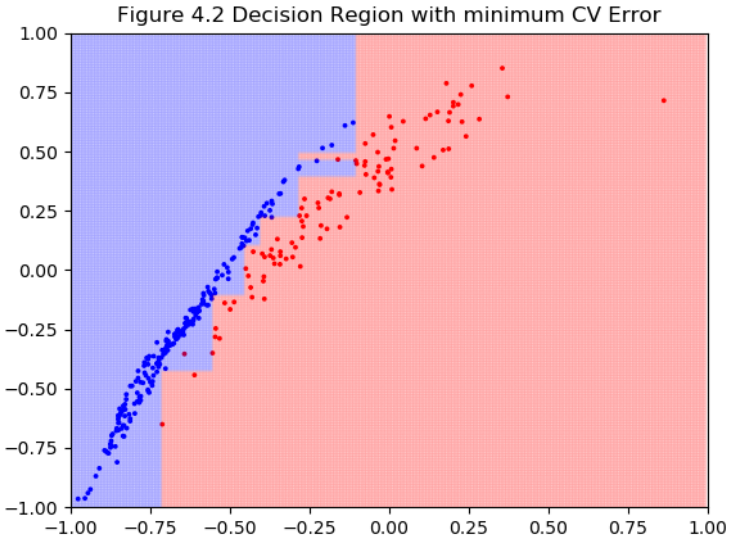
1 Trees/Ensemble Methods

(a)



(b) Looking at the graph, it looks like cv error below max_leaf_nodes = 15 is underfit since the CV error is too high for this trained model.

(c)



(e)

Figure 4.4 CVError for Random Forest with max_leaf_node = 100

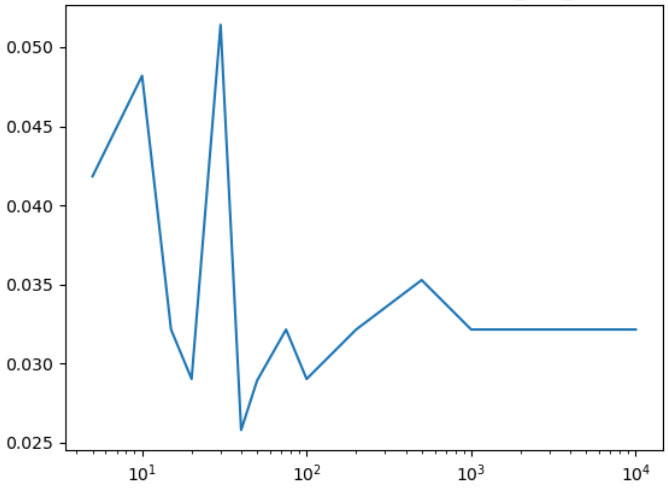


Figure 4.3 CV Error for Random Forest with max_leaf_node = 10

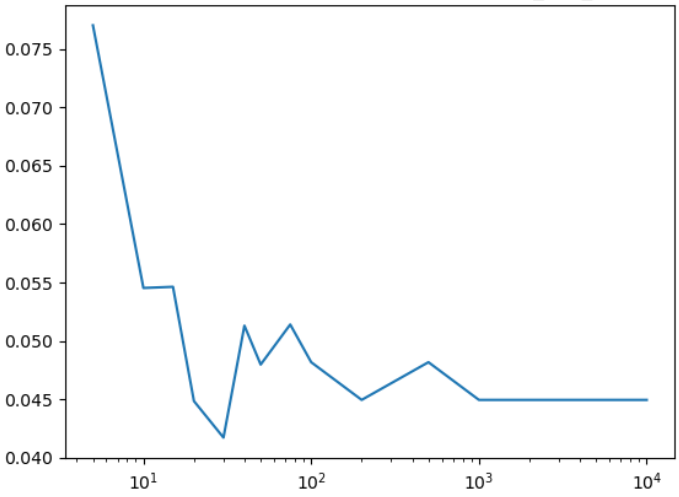
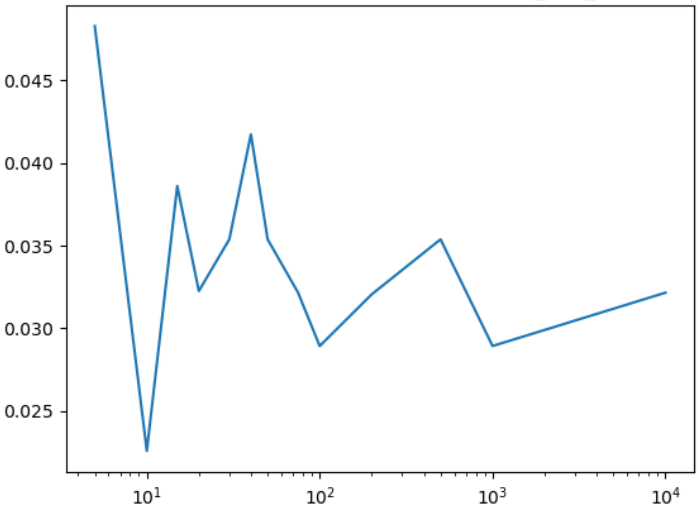


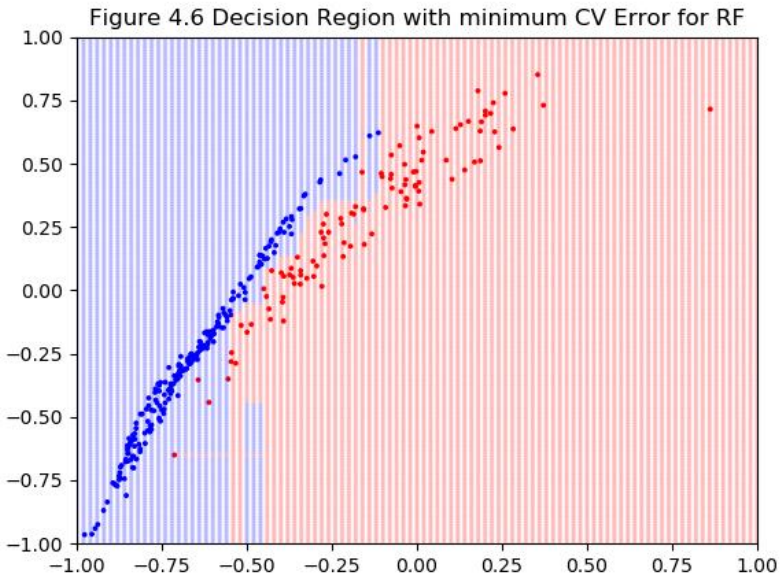
Figure 4.5 CV Error for Random Forest with max_leaf_node = 1000



(f) According to the figure 4.5, max_leaf_nodes = 1000 was impacted the most since using the bagging approach on a well-trained model makes it overfit and hence the error starts going up slowly.

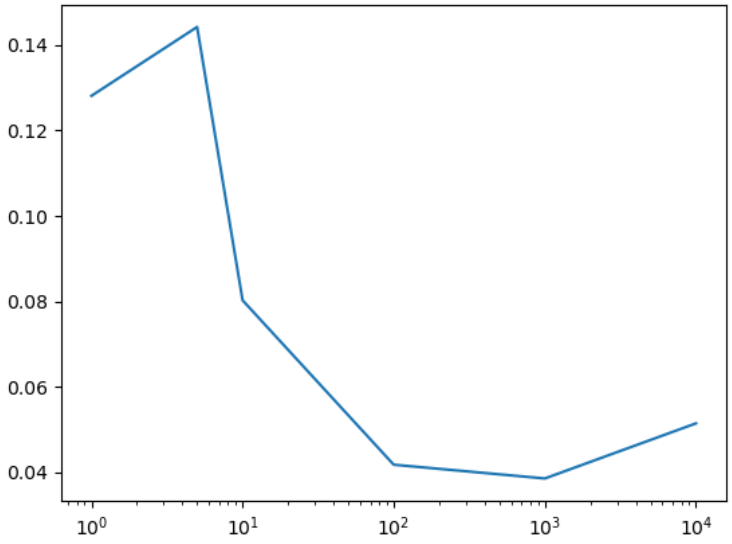
(g) N estimators do make impact on the result; it helps in increasing the accuracy of the model by increasing the number of trees in the forest. But after a point it stops impacting that much, in this case it is n_estimator = 1000. Looking at the figure 4.3,4.4,4.5 we can observe that the error becomes constant even after increasing the n_estimator. This also helps in knowing that now the model is getting overfit.

(h)



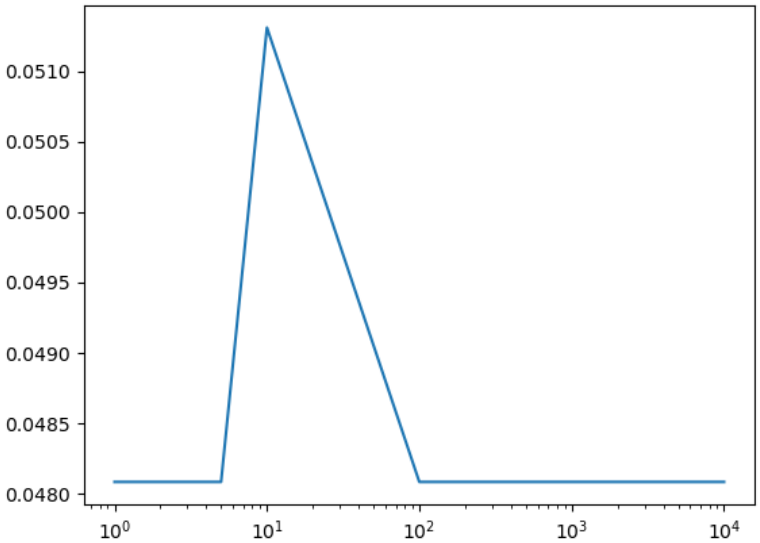
(j)

Figure 4.7 CV Error for AdaBoost max depth =1



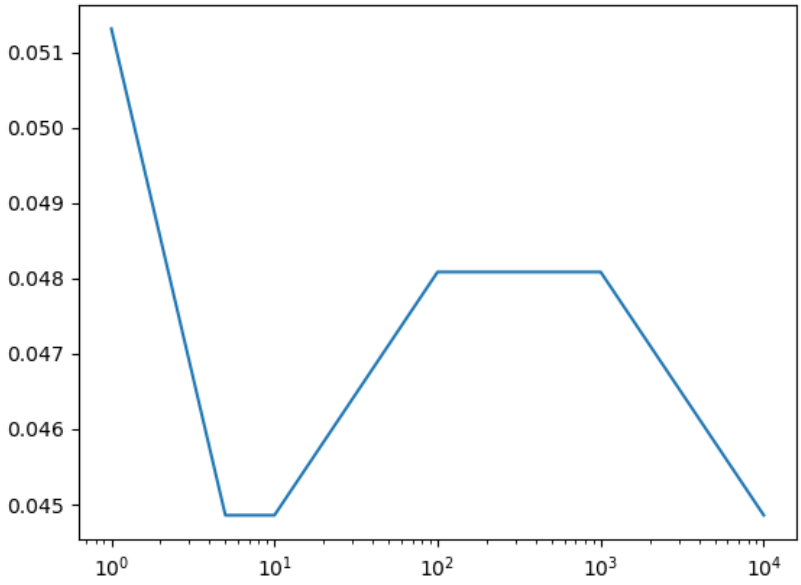
(k)

Figure 4.8 CV Error for AdaBoost with max depth =10



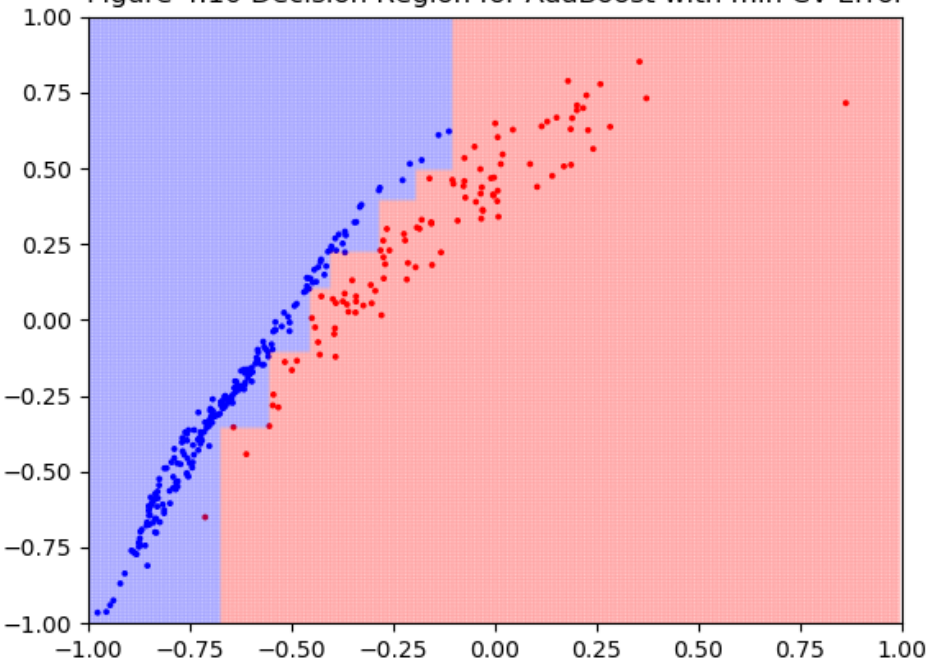
(l)

Figure 4.9 CV Error for AdaBoost with max depth =1000



(m)

Figure 4.10 Decision Region for AdaBoost with min CV Error



2 Reporting Final Error

```
***SVM model***
Bounds for Markov with different confidence interval
99 % -  0.919935948759009
95 % -  0.9839871897518018
75 % -  0.9967974379503604
Bounds for Chebyshev with different confidence interval
99 % -  0.8585220412564268
95 % -  0.9367291333862919
75 % -  0.9717044082512853
Bounds for Hoeffding with different confidence interval
99 % -  0.9539454262246222
95 % -  0.9615717139572929
75 % -  0.9711479191265427
***Neural Network model***
Bounds for Markov with different confidence interval
99 % -  0.7598078462770212
95 % -  0.9519615692554042
75 % -  0.9903923138510808
Bounds for Chebyshev with different confidence interval
99 % -  0.8585220412564268
95 % -  0.9367291333862919
75 % -  0.9717044082512853
Bounds for Hoeffding with different confidence interval
99 % -  0.9539454262246222
95 % -  0.9615717139572929
75 % -  0.9711479191265427
***Random Forest model***
Bounds for Markov with different confidence interval
99 % -  0.7598078462770212
95 % -  0.9519615692554042
75 % -  0.9903923138510808
Bounds for Chebyshev with different confidence interval
99 % -  0.8585220412564268
95 % -  0.9367291333862919
75 % -  0.9717044082512853
Bounds for Hoeffding with different confidence interval
99 % -  0.9539454262246222
95 % -  0.9615717139572929
75 % -  0.9711479191265427
***Adaboost model***
Bounds for Markov with different confidence interval
99 % -  0.5596477181745381
95 % -  0.9119295436349076
75 % -  0.9823859087269815
Bounds for Chebyshev with different confidence interval
99 % -  0.8585220412564268
95 % -  0.9367291333862919
75 % -  0.9717044082512853
Bounds for Hoeffding with different confidence interval
99 % -  0.9539454262246222
95 % -  0.9615717139572929
75 % -  0.9711479191265427
```

(a) According to the values, Adaboost is the model which changes drastically over the confidence interval

(b) I would be choosing Neural Network as my model since it's the most stable ML model in terms of the Final Errors.

(c) The consideration I made for deciding is the sensitivity the model showed for hyperparameters. The model which doesn't drastically changes by slightly tuning the hyperparameters are good models to be trained. As this way the model does not overfit and its easier to find that sweet spot between overfitted and underfitted model.